Introduction to Bioinformatics

Genome Analysis – Sequence contrasts

Overview

One interesting idea that came from considering the review article of Edwards and Rohwer (2005) was that it might be possible to match viral fragments with viral families or host families by some feature inherent in the sequence. The authors suggested looking at GC fraction, codon frequencies, and dinucleotide frequencies. The goal of this set of notes is to test the ability of dinculeotide frequencies to identify the genome of a DNA fragment.

It will be useful to familiarize yourself with a review article that discusses different ways of analyzing DNA fragments and using the analysis to identify pathogenic islands. We won't be devoting a lot of attention to the article below, but there are certain points from it that may help us. Find the following article:

> Samuel Karlin (2001). Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends in Microbiology 9:335-343.

I'll leave it to you to learn from the article what pathogenic islands are. My main interest is the methods put forth in the article to find these blocks of alien genes. Note that the problem posed by Karlin is precisely the opposite as our own. He has in hand a genome and is looking within it for DNA regions that are foreign to it. We have in hand a DNA fragment and are looking for the genome to which it is native. If the method is able to do this, then we might be able to take the DNA fragments that are common results of a metagenomic sequencing project and identify from a set of full microbial and viral sequences the proper homes for the fragments.

Methods to identify anomalous regions (Box 1)

Box 1 of the article puts forth different methods that might be used to identify foreign genes. We'll consider variations on codon usage another time, and amino acid constrasts is a relatively weak method, which we won't consider at all. That leaves two methods: G+C frequency and genome signature contrasts.

Figure 1 shows the results of applying these methods to several bacterial genomes, hoping to find regions within the genomes that display aberrations relative to the whole. It is important to realize that the full genomes are all over a million nucleotides.

SQ1. Given the scale of the x-axes in the graphs shown in Fig. 1 and the average size of a gene (~ 1000 nucleotides), draw to scale a representation of a typical gene on one of the graphs.

The graphs are constructed by considering the genome one window at a time, where a window consists of either 50,000 nt (red lines, I think) or 20,000 nt (black lines, I think). The quantities shown are calculated for the nucleotides within each window, plotted on the graph, and then the window is moved over for the next calculation.

SQ2. I can't find anywhere in the article that gives the identities of the red and black lines. Why do I believe that that the red lines and black lines represent 50,000 nt and 20,000 nt, respectively, and not the reverse?

Comparison of G+C frequencies

Consider the G+C content of the bacterium *Vibrio cholerae*, the causative agent of cholera. Most of the genome has a G+C fraction of about 48%, but there are regions of significant deviation (labeled **A**, **B**, **C**, and **D**). These are considered to be prime candidates for regions of foreign DNA, that is, DNA that was acquired not through the process of cell division but at some point by acquisition from outside the cell, particularly by viral infection.

SQ3. What does G+C fraction mean? Use BioBIKE's GC-FRACTION-OF function to calculate the G+C fraction of the chromosome of A7120 (A7120.chromosome). Then count in the same chromosome the frequency of the four nucleotides and calculate the frequency of each. An easy way to do this is to use COUNTS-OF and *nucleotides* (from the Data menu). DIVIDE the result by the LENGTH-OF the chromosome. What is the relationship between the four frequencies thus derived and the GC-fraction?

Our plan is to use GC-fraction, if possible, to identify the virus to which a read-sized DNA fragment belongs. This is analogous to identifying a person's country by the person's height. It works if everyone in a country has the same height – everyone in Sweden is 6'2" and everyone in Burma is 5'6". Now suppose we encounter a person who tells us he's from Burma or Sweden – he forgets which. We note that he's 5'6". Shall we inform him that he's from Burma? We can, if everyone in a country is the average height, plus or minus an inch or so. But what if there were a significant number of 5'6" people in Sweden? Our method of classification fails if the variation of heights ***within*** a country leads to significant overlap in heights of people ***between*** countries. It is therefore important to know not only the average G+C fraction but also the distribution within an organism and how that compares to the range of G+C fractions amongst different organisms.

SQ4. What is the range of overall G+C fractions in different cyanobacteria? Having developed a method to get the G+C fraction for a specific organism, you're in a good position to get it for all organisms, by generalizing the form you made in SQ3. If you replace `A7120` with `organism`, you can think of the form as a general function:

_f_(organism) = [form from SQ3 using organism as a variable]

Now paste that function into the body of APPLY-FUNCTION-OF and apply it to *all-cyanobacteria* (obtainable from the DATA menu).

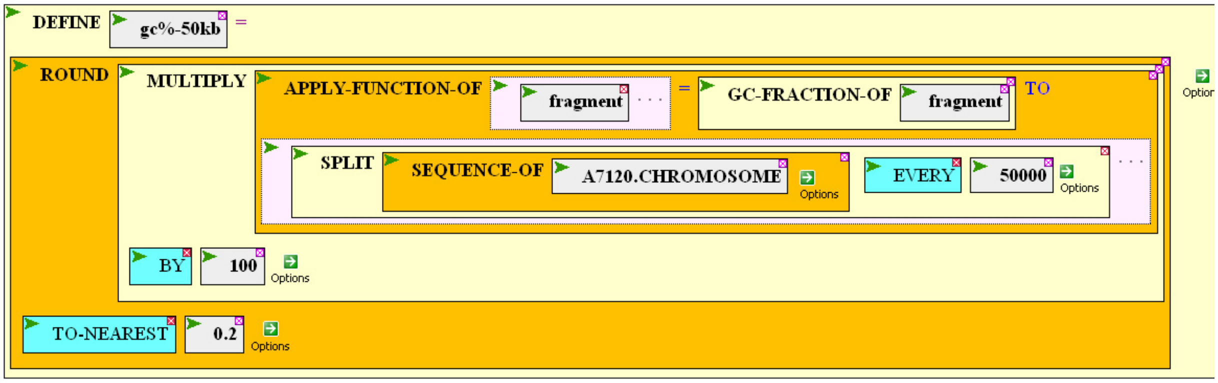
(To tell the truth, you could have done the same thing more simply using implicit mapping, but this is practice!). That's a pretty wide range – the range amongst all vertebrates, for example, is far less. What about the GC-fractions of fragments taken from a single organism? My strategy is to SPLIT the chromosome of A7120 into fragments, every 50,000 nt (as in Fig. 1 of the article – remember the article?), and then apply GC-FRACTION-OF.

SQ5. What is the range of G+C fractions in 50-Kb (kilobase) fragments of A7120? You're already set up to do this – just modify your code for SQ4. Define a variable (e.g. GC%-50kb that contains the result of GC-FRACTION-OF.

The problem this time is in the interpretation, as there are too many numbers to eyeball. One thing that would help is to simplify the numbers to percentages. Of course, you can convert the fractions to percents by MULTIPLYing them by 100, but that just gives you very long percents. What you want to do is to ROUND the percents to the nearest 0.2.[*] Let's do it.

[*] Why 0.2%, not 0.1%? Because the smallest percent difference in 500 nt is 1/500 = 0.2%.

SQ6. Re-DEFINE the variable of SQ5 so that it contains the range of G+C fractions in 50-Kb (kilobase) fragments of A7120 as percents, showing only one decimal place. You can do this using the form shown below.



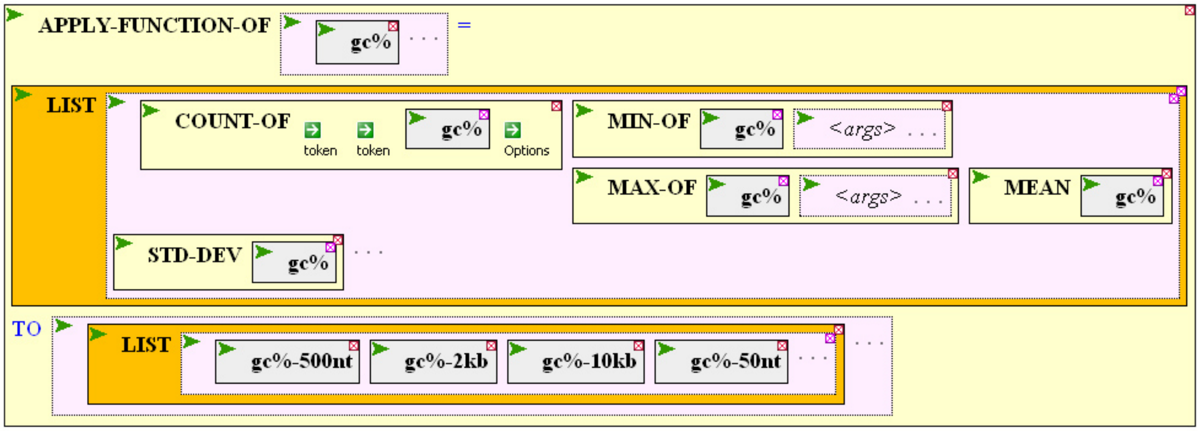
```
DEFINE gc%-50kb =
  ROUND
    MULTIPLY
      APPLY-FUNCTION-OF fragment ... = GC-FRACTION-OF fragment TO
        SPLIT SEQUENCE-OF A7120.CHROMOSOME Options EVERY 50000 Options ...
      BY 100 Options
    TO-NEAREST 0.2 Options
```

We still haven't determined the range of this list of G+C fractions. You can do this by using the MIN-OF and MAX-OF functions (Arithmetic menu, Aggregate arithmetic) and get additional statistical information by using the MEAN and STD-DEV functions (Arithmetic menu, Statistics). And you'll want to know how many fragments you've analyzed, so add COUNT-OF. It will be convenient (as you will see in a moment) to perform all five operations at once, within a LIST.

SQ7. Calculate the number of fractions, the minimum value, the maximum value, the mean, and the standard deviation of the list of G+C fractions expressed as percentages with one decimal place, all within a single list. Do this by bringing down LIST from the List menu, adding five holes, and filling the holes with the desired quantities.

You've just reproduced Karlin's G+C analysis on *Anabaena*. But we want to know the G+C fractions of read-sized fragments, not 50,000 nt but 500 nt. How does reducing the size of the fragments affect the quantities you calculated? Obviously, the smaller the size of the fragment, the greater the number of fragments. What of the other quantities?

SQ8. Repeat SQ6 and SQ7 for fragments of sizes 10,000, 2,000, and 500 nt. This should be almost trivial, requiring you only to change the fragment size and make up an appropriate name for the defined variable. You can make your life easier calculating the statistics (SQ7) by generalizing the calculation and applying the generalized function of some variable to a list consisting of GC%-500nt, GC%-2kb, GC%-10kb, and GC%50kb. As below:



```
APPLY-FUNCTION-OF gc% ... =
  LIST
    COUNT-OF token token gc% Options
    MIN-OF gc% <args> ...
    MAX-OF gc% <args> ...
    MEAN gc%
    STD-DEV gc% ...
TO
  LIST gc%-500nt gc%-2kb gc%-10kb gc%-50nt ... ...
```

When you execute this form, the statistics will appear in the results section, but it will take some imagination to compare them with one another. It would be much nicer to have them nice and neat in a table. BioBIKE has limited graphical capabilities, so we outsource. The solution is to bring the results into Excel (or equivalent) and let it make the table. Excel can read tab-delimited files (translating numbers separated by tabs into numbers placed in separate columns), so that's what we'll give it.

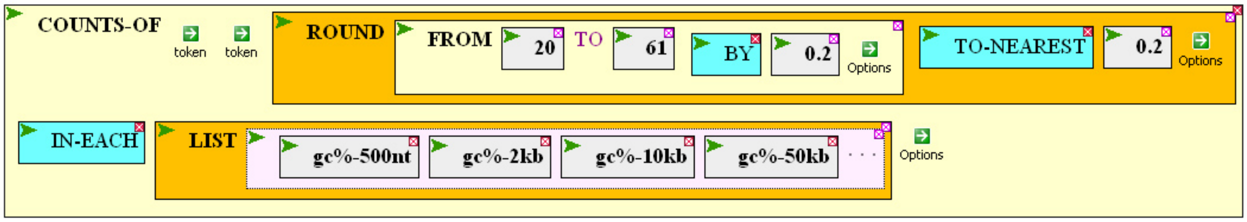
SQ9. Write the results of SQ8 to a tab-delimited file, using the WRITE function (Input-output menu), pasting the form from SQ8 into the FROM field and making up any name you like for the file (so long as it is within quotes… and it will be easier in the end if you give it a .txt extension).

SQ10. Find the file you just wrote, by going to the File menu (of BioBIKE, not your browser), clicking Files, and clicking on the file you just created. Download the file to your own computer by right-clicking the filename and saving it somewhere.

SQ11. Bring the file you just saved into Excel, as a tab-delimited file. Add a first row and a first column and add appropriate labels.

SQ12. What is the effect of reducing fragment size on the range of G+C fractions?

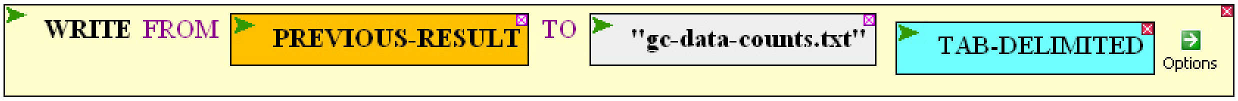
The numbers you've gotten from this analysis might provide you with some insight into the distribution of G+C fractions in progressively smaller fragments of DNA. But it's all abstract. Nothing beats a picture. Again we give the graphical job to Excel (or equivalent). You can save the data in tab-delimited format and use Excel magic to make a binned frequency curve, but if you don't know that magic, BioBIKE magic is probably simpler. The goal is to count how many times each number occurs. COUNTS-OF can do the whole thing in one gulp:



```
COUNTS-OF  token token  ROUND  FROM 20 TO 61 BY 0.2 Options  TO-NEAREST 0.2 Options
IN-EACH  LIST  gc%-500nt gc%-2kb gc%-10kb gc%-50kb . . .  Options
```

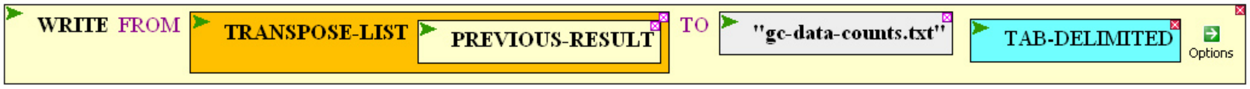
[*You no doubt did not think of using ROUND, but if you don't, the function will find no matches (except for 20). This is because you're comparing numbers, and a machine that thinks in base 2 does not consider 20 + .2 + .2 + .2 + .2 + .2 exactly equal to 21, just as we don't consider 20 + .333 + .333 + .333 exactly equal to 21. Most decimals that terminate in base 10 go on forever in base 2, just as 1/3 goes on forever in base 10.*]

Now to save the file in tab-delimited form, so that Excel can read it. Doing this directly:



```
WRITE FROM PREVIOUS-RESULT TO "gc-data-counts.txt" TAB-DELIMITED Options
```

will produce a tab-delimited file where data from gc%-500nt occupies the first row and data from gc%-50kb occupies the fourth row. Unfortunately, Excel isn't equipped to handle rows with more than 256 columns, and there are far more 500-nt fragments than this. However, Excel ***can*** handle up to 65,536 rows. So all we need to do is to make the rows columns and the columns rows. This process is called transposition (one of many meanings of the term), and you can achieve it by using the TRANSPOSE-LIST function (List-tables menu, List production):



```
WRITE FROM TRANSPOSE-LIST PREVIOUS-RESULT TO "gc-data-counts.txt" TAB-DELIMITED
```

SQ13. Write to your personal BioBIKE directory a tab-delimited file consisting of GC-fraction counts of the four results you've obtained. Download the file to your computer and upload it into Excel, as described in SQ10 and SQ11. Insert a first row and label each of the four columns.

SQ14. Create a line chart of the results in Excel.

SQ15. What is the effect of reducing fragment size on the range of G+C fractions?

SQ16. Considering (amongst other things) the range of GC-fractions in *organisms* (e.g. all cyanobacteria), what limitations are there on using GC-fraction in viral fragments to identify their parent virus or host organism?

Coming Attractions!

Sequence contrasts using dinucleotides

Discussion of Karlin (2001) Box 2 How to do it in BioBIKE.

Codon usage contrasts

Discussion of Karlin (2001) Box 3. How to do it in BioBIKE.

Tests of significance

How to tell when two similar profiles (GC%, dinucleotide frequency, codon usage) are close enough to draw a conclusion?

Doing this and more with unknown viral sequences!